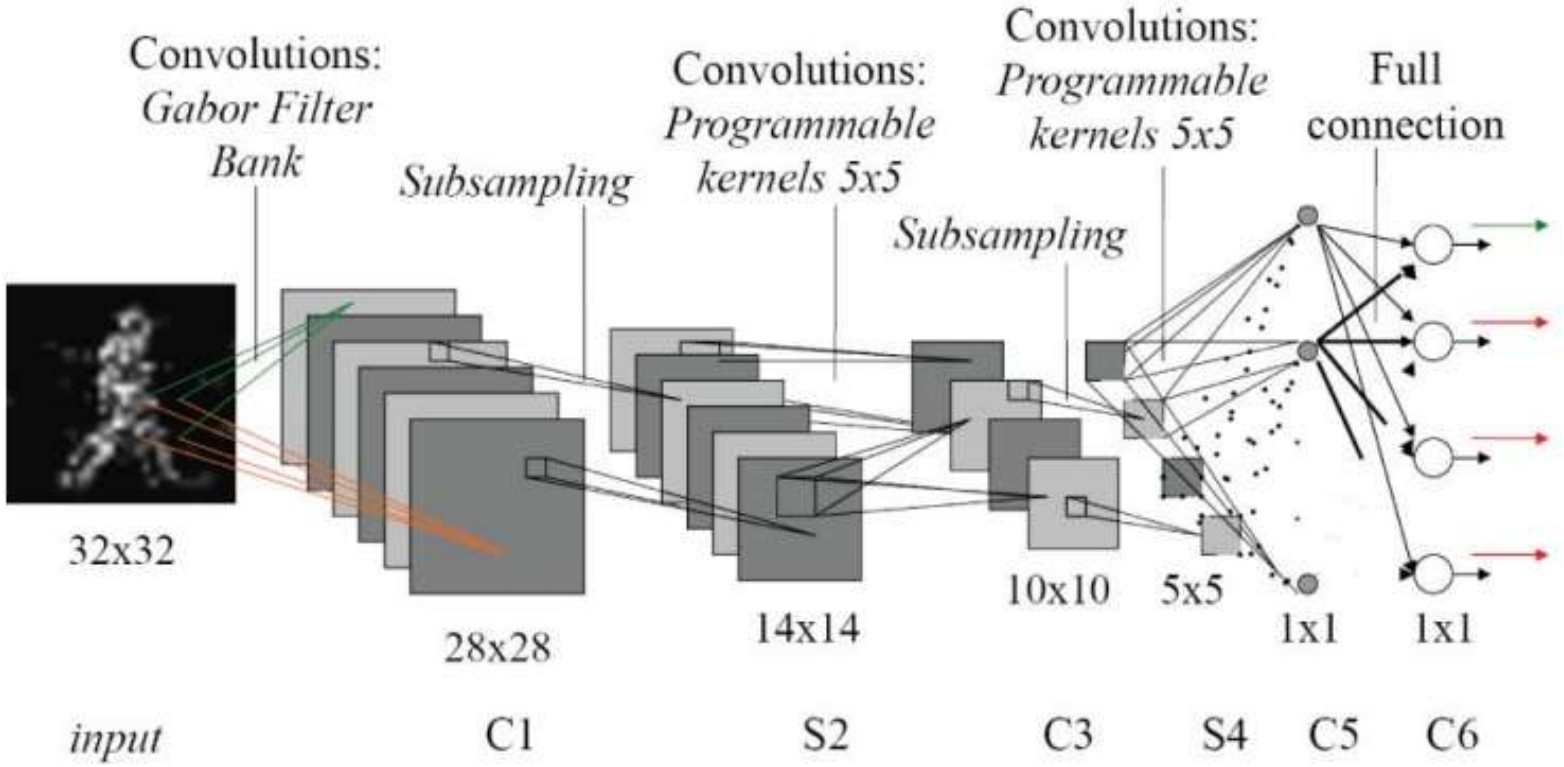
Spiking row-by-row FPGA Multi-kernel and Multi-layer Convolution Processor.

Ricardo Tapiador Morales
Robotic & Tech of Computers Lab,
University of Seville
ricardo@atc.us.es

Convolutional neural networks

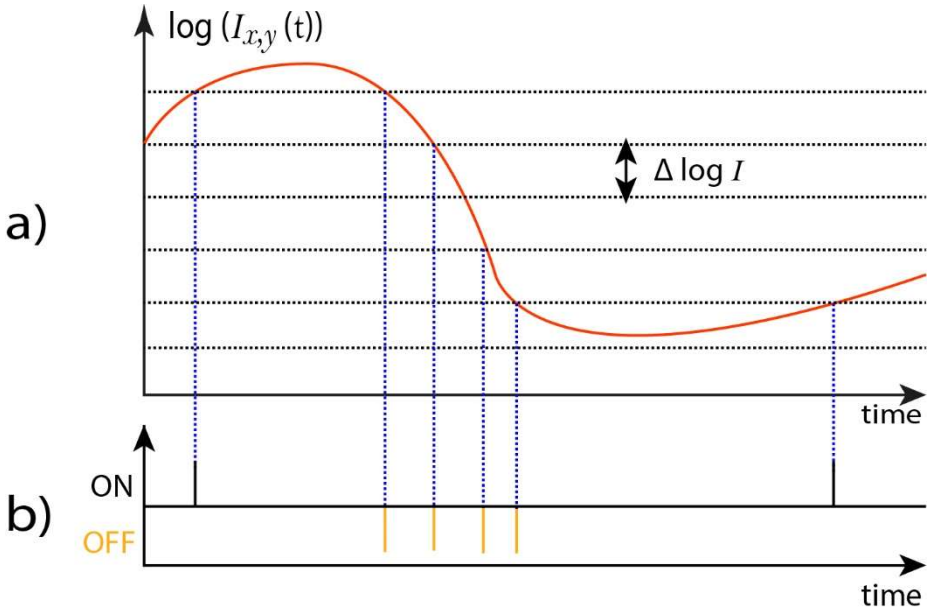


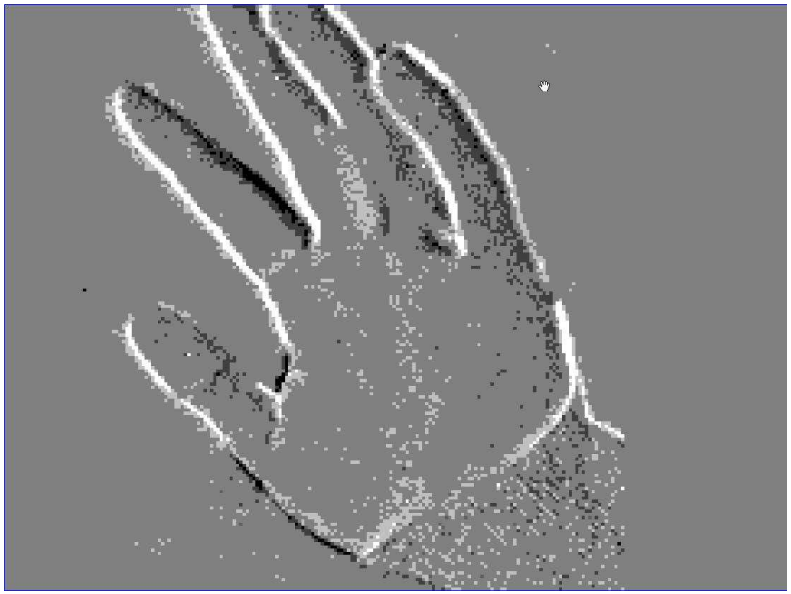
Neuromorphic Engineering

- Neuromorphic engineering mimics the behavior of the human brain, where information is encoded in spikes (also called events) that are processed in parallel by massive layers of neurons interconnected via synapses.

Dynamic Vision Sensors

Dynamic Vision Sensors



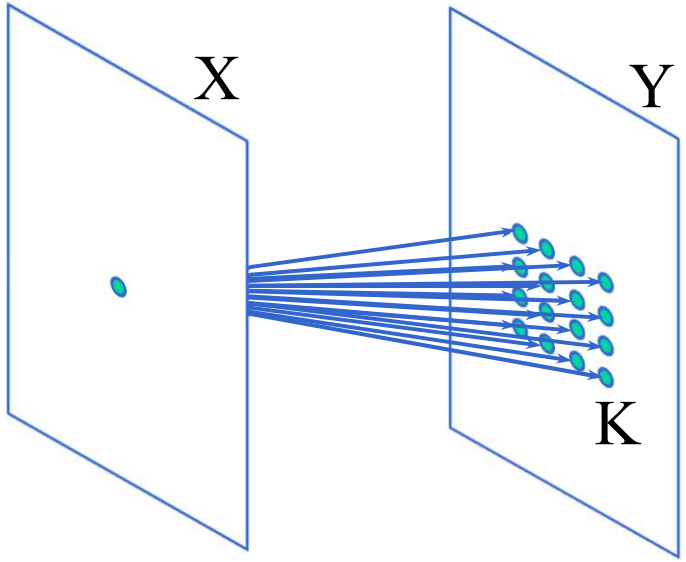


Convolution with Spikes

$$\forall_{i,j} \rightarrow Y(i,j) = \sum_{a=-n/2}^{n/2} \sum_{b=-m/2}^{m/2} K(a,b) \cdot X(a+i, b+j)$$

Y: convolution result. *K: nxm* kernel matrix *X:* input image

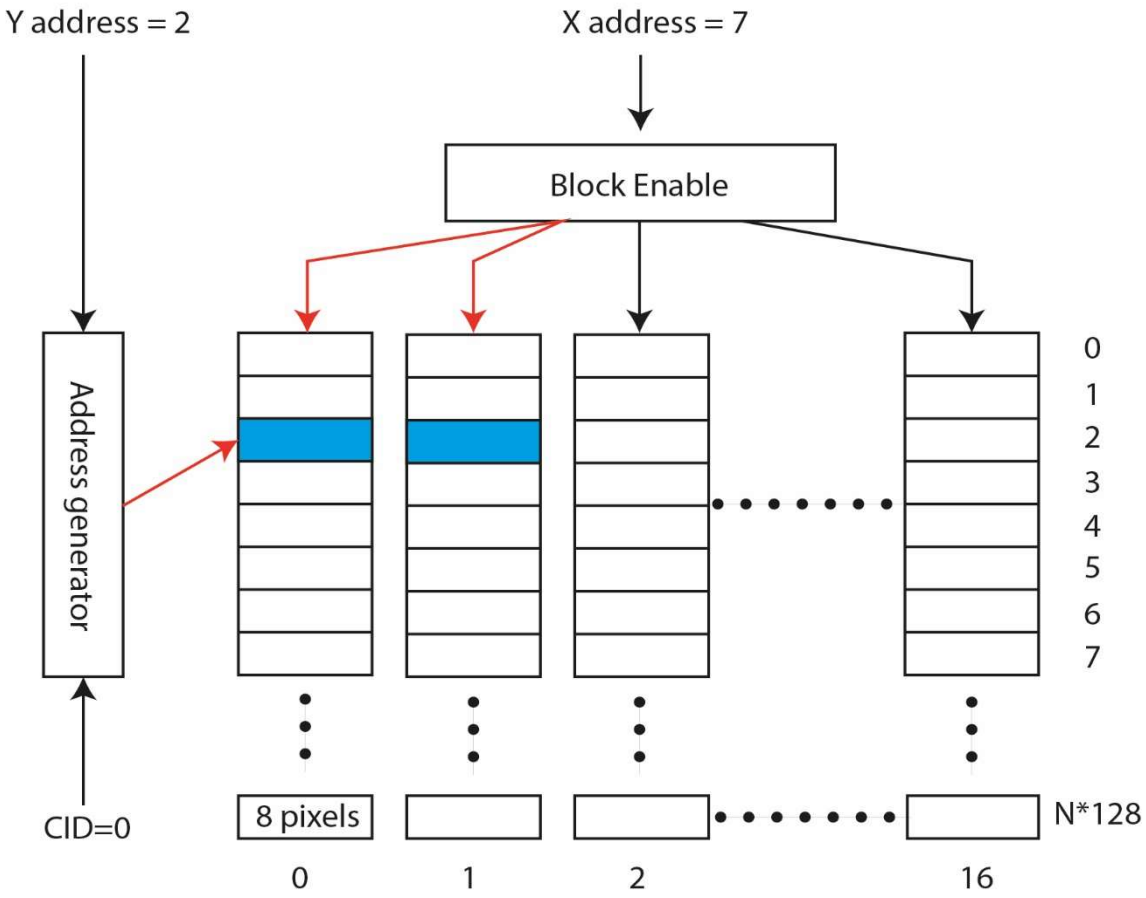
- X(i,j) data can be coded in frequency of events.
- Each event implies to accumulate K into the Y neighborhood around Y(i,j).



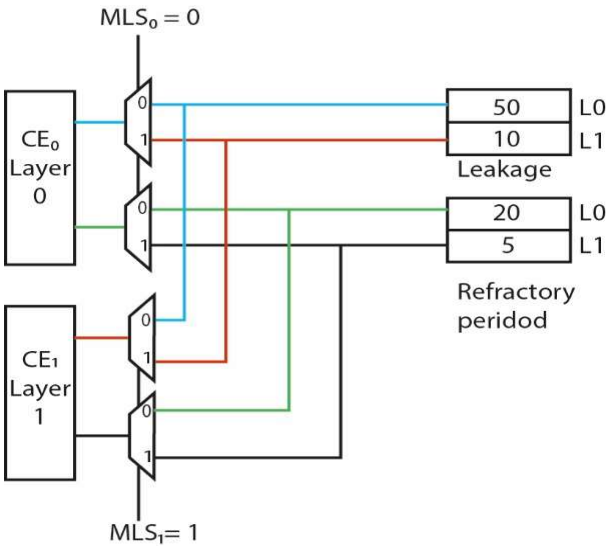
- Y output is based on LIF neuron.
- X, K, Y allows signed values.
- Each (i,j) event implies:

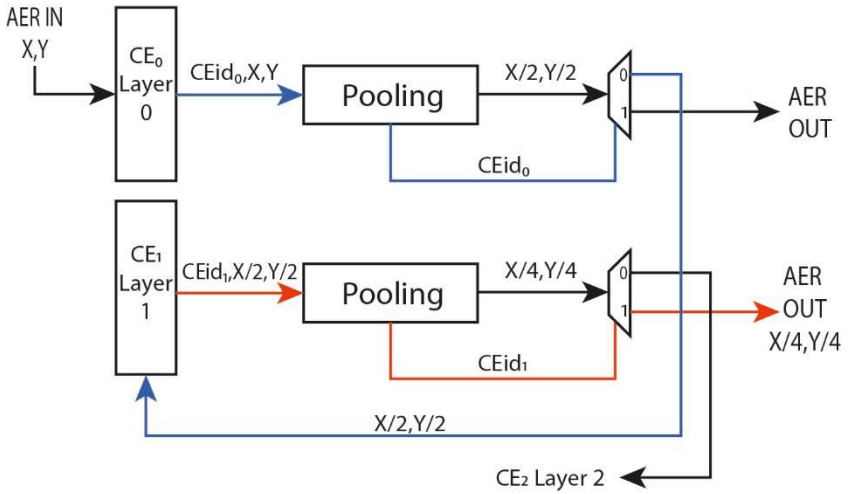
$$Y(i+a, j+b) = Y(i+a, j+b) + K(a,b), \quad \forall a,b \in \dim(K)$$

Convolution processor architecture: Memory

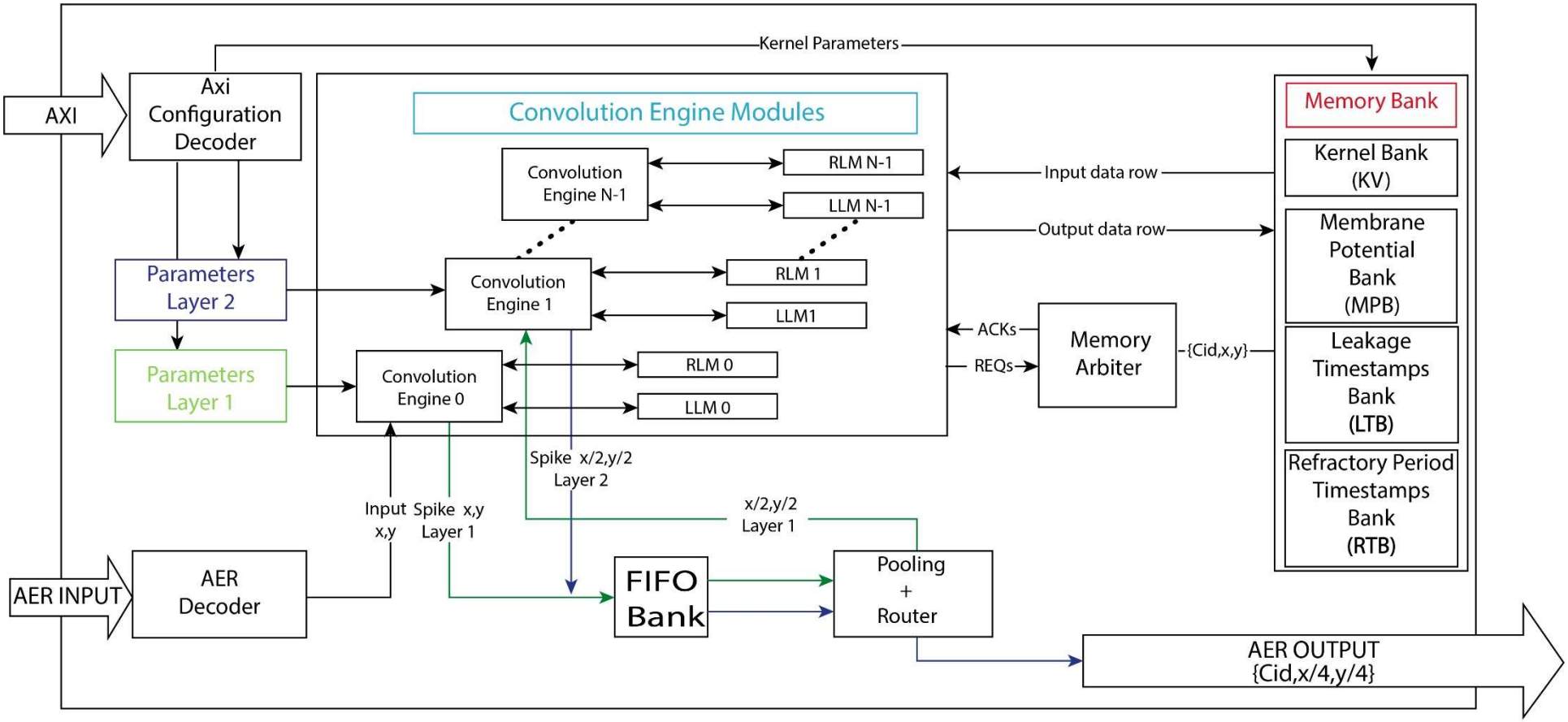


Convolution processor architecture: Multi-kernel and Multi-layer mechanisms

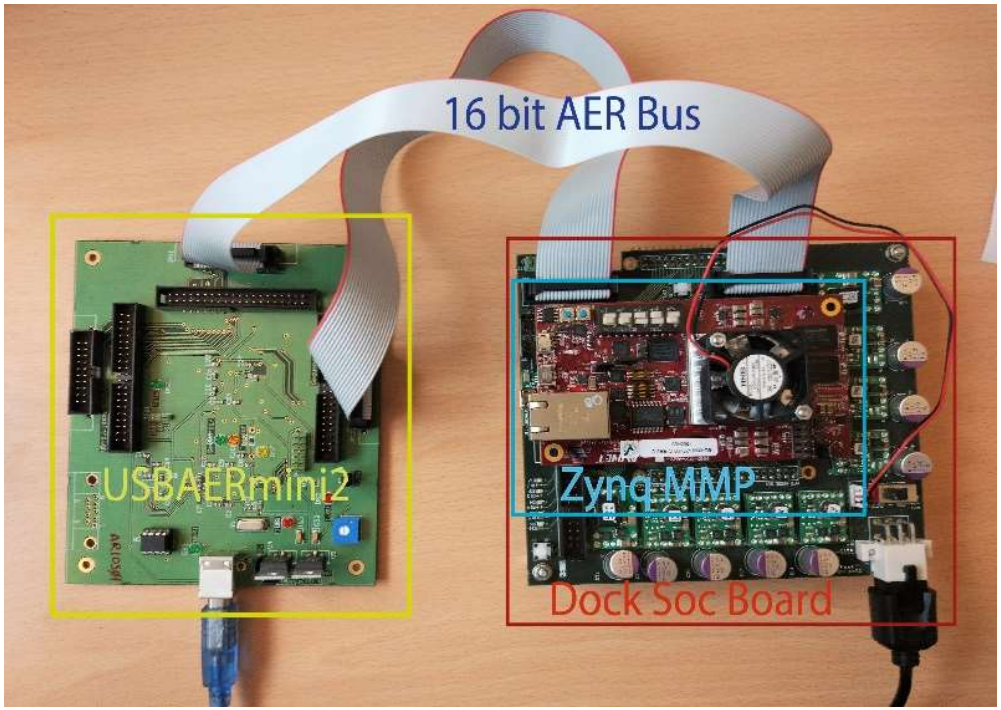


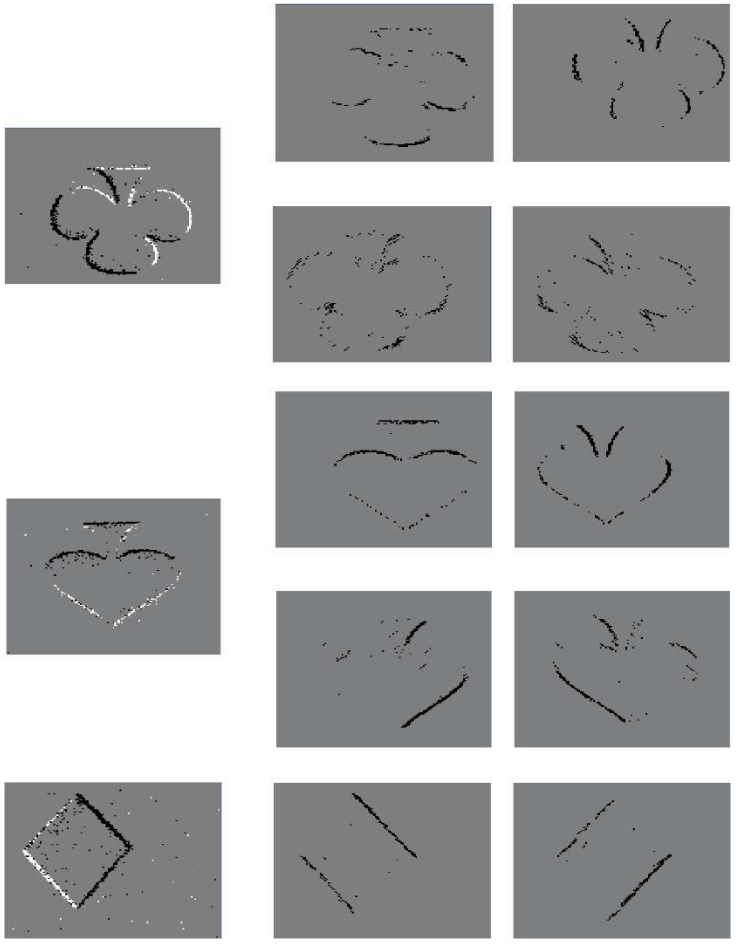


Convolution processor architecture



Test Scenario





Results:

Area

Resource	Utilization	Available	Utilization %
LUT	257503	277400	92,83
LUTRAM	50851	108200	47.00
FF	179925	554800	32,43
BRAM	713.5	755	94,37
IO	43	362	11,88

Performance

- Latency : 1.44-9.98 µs
- Input Throughput : 0.10-0.69 Meps

Future Works

- Implement a Spiking Convolutional Neural Network

- Add mechanisms to auto-configure the different parameters

Thanks for your attention